Abstract Title Page

Title:

Test Format and the Variation of Gender Achievement Gaps within the United States

Authors and Affiliations:

Sean Reardon, Stanford University
Erin Fahle, Stanford University
Demetra Kalogrides, Stanford University
Anne Podolsky, Learning Policy Institute
Rosalia Zarate, Stanford University

Abstract Body

Background / Context:

Prior research demonstrates the existence of gender achievement gaps and the variation in the magnitude of these gaps across states (Reardon et al., forthcoming; Pope & Snydor, 2010; Guiso et al. 2008, Hyde et al., 2008). Notably, research has shown that the size and direction of the average gender gap varies by subject. In mathematics, research shows that, on average, gender achievement gaps on mathematics state standardized tests are small in magnitude and do not consistently favor males or females across the states (Fryer & Levitt, 2009; Lee, Moon, & Hegar, 2011; Robinson & Lubienski, 2011; Penner & Paret, 2008; Sohn, 2012). On the other hand, in reading, a significantly larger consistently female-favoring gap exists, which varies in magnitude across the states (Chatterji, 2006; Fryer & Levitt, 2009; Husain & Millimet, 2009; Robinson & Lubienski, 2011).

A smaller body of research reveals that test format, the proportion of multiple-choice or constructed-response questions, is a key influence on male and female students' performance (Lindberg et al., 2010; Beller & Gafni, 2000; Gamer & Engelhard, 1999; DeMars, 1998; Ben-Shakhar & Sinai, 1991). The general findings from the literature are that males perform better than females on multiple-choice questions, and that females perform better on constructed-response test questions. This prior research, however, has focused on a narrow set of tests and does not characterize the extent to which the different state accountability test formats explain the variation in the magnitude of achievement gaps across states. A large-scale, systematic analysis has yet to be undertaken.

Purpose / Objective / Research Question / Focus of Study:

In this paper, we characterize the extent to which the variation of gender achievement gaps on standardized tests across the United States can be explained by differing state accountability test formats. A comprehensive analysis of the interplay between state standardized test formats and differences in gender achievement on those tests is important for informing policies and practices that aim for greater equity in education. Specifically, these tests are increasingly used to label, and often reward, states. As a result, it is critical that we better understand how the selection and use of different accountability assessments, with different test formats, may distort the interpretation of the size and direction of the gap, impacting comparisons across states.

Setting:

This study focuses on gender achievement gaps in the United States, with particular attention to explaining the state-level differences in the levels of the gaps.

Population / Participants / Subjects:

This study performs both a state-level and district-level analysis. We use student test score results in grades 4 and 8 in 2009 from three different tests: (1) state accountability tests (we have data from all 50 states and roughly 9,400 school districts); (2) the Measures of Academic Progress (MAP) assessment administered by the Northwest Evaluation Association (NWEA) (we

have data from roughly 3,700 school districts); and (3) NAEP tests administered (all 50 states). State accountability tests vary in item format among states; the NWEA and NAEP tests have a common item structure across states.

For the state-level analyses, the population of study is the set of 48 U.S. states in the 2008-09 school year. Specifically, we analyze the subject-specific gender achievement gaps on three different tests among fourth and eighth grade students in each state. We use the National Center for Education Statistics' EdFacts Database and NAEP assessment data to explore average state-level gender achievement gaps in fourth and eighth grade in ELA and mathematics for the 2008-09 school year. For the district-level analyses, the population of study is the set of districts in both the NWEA and EdFacts data, which is approximately 700 districts. Again, we analyze the gender achievement gaps in the 2008-09 school year among fourth and eighth grade students.

Intervention / Program / Practice:

In this study, we evaluate the impact of test item format on gender differences in achievement on state assessments. As noted above, female students tend to perform less well on multiple choice questions (Lindberg et al., 2010; Beller & Gafni, 2000; Gamer & Engelhard, 1999; DeMars, 1998; Ben-Shakhar & Sinai, 1991), which has important implications for how the format of a test could induce variation in the scores that does not correspond to underlying ability differences between the sexes (note that we make no assumption about the underlying differences, but rather focus on how the test format can change the observed gender achievement gap). In our setting, we focus specifically on how differences in test format on state assessments can explain observed differences in the achievement gap across states or districts on these assessments, as compared with a uniform assessment.

Research Design:

In order to understand the effects of test item format on gendered achievement, we leverage the variation in the item format across three assessments to model how within-state gender achievement gaps vary with differences in test item format, specifically the proportion of multiple-choice, short constructed-response, and extended-response questions. The state assessments, captured in the EdFacts data, vary in test format across states and districts in different states; whereas, the NAEP and NWEA MAP assessments have a standard format across all states and districts.

Data Collection and Analysis:

We estimate achievement gaps in each state or district using the V-statistic (Ho, 2009; Ho & Haertel, 2006; Ho & Reardon, 2012). First, we estimate gender gaps in mathematics and reading and in grades 4 and 8 within each state from both the 2009 state accountability tests (EdFacts data) and the 2009 NAEP tests. Then, we fit a model that the gender achievement gaps on among students in state or district s on the two different tests. Denote the state accountability test (whose format varies among states) as a, and, the national test (which is identical in each state/district) as n. For each test let p_{st} be equal to the proportion of non-multiple choice question (note that $p_{sn} = p_n$ is a constant, since test n is identical in each state/district, but p_{sa} varies across states). Now suppose we estimate G_{st} with error w_{st}, where $w_{st} \sim N(0, \omega_{st}^2)$, so that $\hat{G}_{st} = G_{st} + w_{st}$. Let T_{st} be an indicator variable for the state accountability test (so $T_{sa} = 1$ and

$T_{sn} = 0$), and define $\alpha = -\delta p_n$. Then we can express the estimated gap on test t in state or district s as:

$$\begin{aligned}\hat{G}_{st} &= \gamma_s + \delta p_{st} + v_{st} + w_{st} \\ &= \gamma_s + (\delta p_{sa})T_{st} + (\delta p_n)(1 - T_{st}) + v_{st} + w_{st} \\ &= -\alpha + \gamma_s + (\delta p_{sa} + \alpha)T_{st} + v_{st} + w_{st} \\ &= -\alpha + \gamma_s + \alpha T_{st} + \delta(p_{sa} \cdot T_{st}) + v_{st} + w_{st}\end{aligned}$$

We can estimate α and δ by fitting model (3) using a precision-weighted fixed-effects regression model, including state or district fixed effects and weighting each observation by $\hat{\omega}_{st}^{-2}$, the inverse of the estimated sampling variance of $\hat{G}_{st}$. The use of state or district fixed effects in the models means that we are essentially controlling for the true gender gap in each state in the models. We fit multiple versions of this model: (1) separately for each grade and subject combination, (2) pooling across grades within subjects, (3) pooling across subjects within grades, and (4) pooling across all four grade-subject combinations. In these latter models (2) – (4), we include state-by-subject, state-by-grade, or state-by-grade-by-subject fixed effects as appropriate.

We repeat this process at the district level, starting with estimating the gender gaps in grades 4 and 8 in mathematics and reading in each of the school districts for which we have data available from both the EdFacts and NWEA tests. Again, because the NWEA tests have a common item structure across states (and therefore districts), they operate in the model in the same way as the NAEP tests do in the state tests. We fit a similar set of models at the district-level: a model separately for each grade and subject combination, and models pooling across grades within subjects, across subjects within grades, and across all four grade-subject combinations. In these latter models we include district-by-subject, district-by-grade, or district-by-grade-by-subject fixed effects as appropriate. Because there are multiple observations per state, we cluster the standard errors in these models at the state level.

Findings / Results:

We first note that states vary substantially in the proportion of multiple-choice items on their tests in mathematics and ELA (ranging from 50-100% multiple-choice). We find that boys do better on multiple-choice tests than girls of the same academic skill. Specifically, our estimates imply that gender gaps are, on average 0.22 standard deviations greater (favoring boys more) on multiple-choice tests than on constructed response item tests. These results appear to be driven primarily by gender-by-item format interactions affecting performance on ELA tests: in ELA, gender gaps on multiple-choice tests are roughly 0.30 to 0.40 SD larger (favoring girls less and boys more) than on constructed- response tests. On mathematics tests, the difference in performance is roughly 0.10 SD smaller, but still favoring girls less and boys more on multiple-choice tests than on constructed-response tests. (please insert table 1 here). These patterns are consistent regardless of whether we use NAEP or NWEA tests as the audit test.

Conclusions:

This is the first analysis of its kind to explore the interplay between each state's standardized test format and differences in gender achievement on those tests. Such an understanding is timely because scores on these tests may have consequences for schools and students. Moreover this paper provides a better understanding of how the selection and use of

different accountability assessments, with different test formats, may distort the interpretation of the size and direction of the gap, impacting comparisons across states.

Appendices

Appendix A. References

Andreescu, T., Gallian, J. A., Kane, J. M., & Mertz, J. E. (2008). Cross-Cultural Analysis of Students with Exceptional Talent in Mathematical Problem Solving. *Notices of the AMS*, *55*(10), 1248–1260.

Beller, M., & Gafni, N. (2000). Can Item Format (multiple choice vs. open-ended) Account for Gender Differences in Mathematics Achievement?. *Sex Roles, 42*(1-2), 1-21.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender Differences in Multiple-Choice Tests: the Role of Differential Guessing Tendencies. *Journal of Educational Measurement*, *28*(1), 23-35.

Chatterji, M. (2006). Reading Achievement Gaps, Correlates, and Moderators of Early Reading Achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*(3), 489.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279-299.

Fryer Jr, R. G., & Levitt, S. D. (2009). *An empirical analysis of the gender gap in mathematics* (No. w15430). National Bureau of Economic Research.

Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29-51.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science 320*(5880), 1164.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201-228.

Ho, A. D., & Haertel, E. H. (2006). Metric-free measures of test score trends and gaps with policy relevant examples (CSE Report No. 665). Los Angeles, CA: *Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies.*

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489-517.

Husain, M., & Millimet, D. L. (2009). The mythical "boy crisis"? *Economics of Education Review*, (28), 38–48. doi:10.1016

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494-495.

Lee, J., Moon, S., & Hegar, R. L. (2011). Mathematics skills in early childhood: Exploring gender and ethnic patterns. *Child Indicators Research*, *4*(3), 353-368.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological Bulletin*, *136*(6), 1123.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, (37), 239–253.

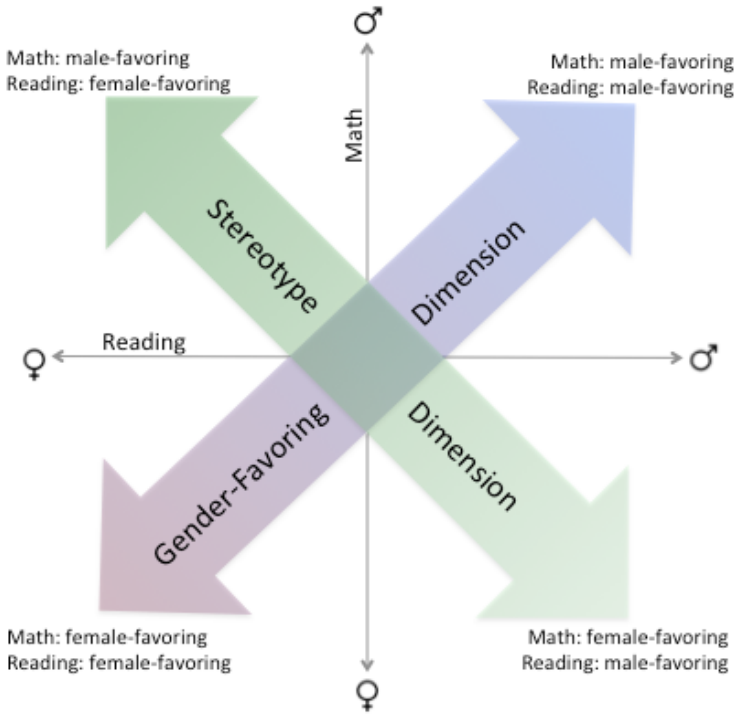
Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *The Journal of Economic Perspectives*, *24*(2), 95-108.

Robinson, J. P., & Lubienski, S. T. (2011). The Development of Gender Achievement gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings. *American Educational Research Journal*, *48*(2), 268–302.

Sohn, K. (2012). A new insight into the gender gap in math. *Bulletin of Economic Research, 64*(1), 135-155.

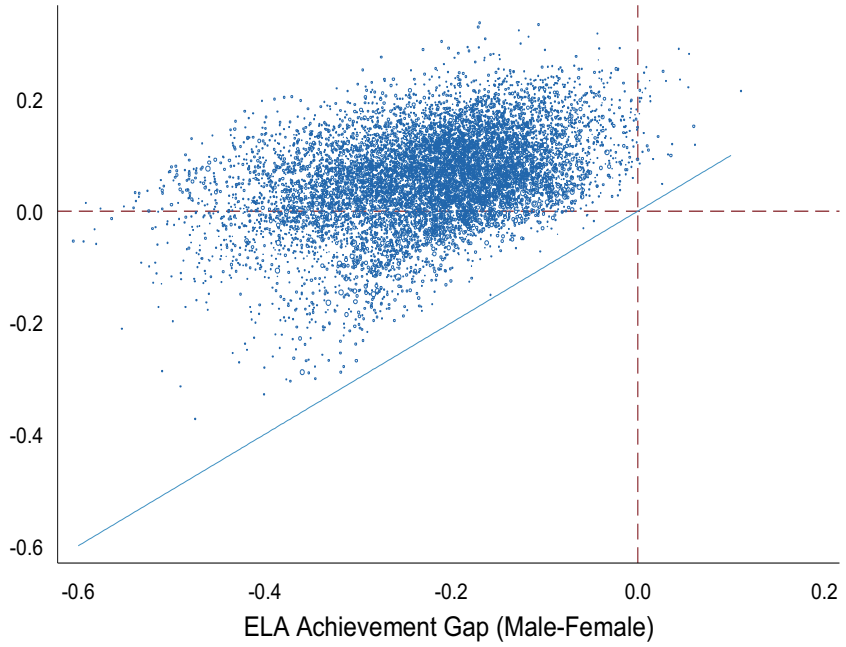
Appendix B. Tables and Figures

Figure 1: Dimensions of Gender Achievement Gaps



Notes: Reading gaps are plotted on the x-axis and corresponding mathematics gaps on the y-axis. Positive (negative) values indicate gaps are male-favoring (female-favoring). Gender equality in the gaps is at the origin.

Figure 2: Male-Female Mathematics and ELA Achievement Gaps, School Districts 2009-2012



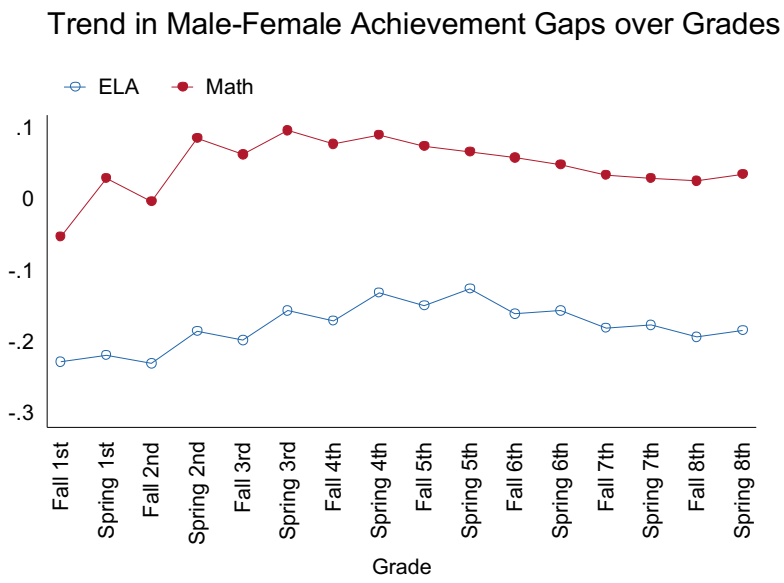
Notes: Reading gaps are plotted on the x-axis and corresponding mathematics gaps on the y-axis. Positive (negative) values indicate gaps are male-favoring (female-favoring). The model used to estimate average gaps includes state fixed effects and adds the average state NAEP gap to the Empirical Bayes estimate.

Table 1: Relationship between Proportion Multiple-Choice Items on State Tests and the Size of Gender Gaps, State-Level

	Mathematics				ELA			
	Grade 4		Grade 8		Grade 4		Grade 8	
Model 1: State-Level NAEP Audit Test								
Proportion Short Response+	-0.135	**	-0.109		-0.223	*	-0.376	**
Extended Response	(0.041)		(0.075)		(0.084)		(0.113)	
Model 2: District-Level NWEA Audit Test								
Proportion Short Response+	-0.126		-0.151	*	-0.296	**	-0.389	***
Extended Response	(0.122)		(0.068)		(0.098)		(0.101)	

All models are weighted by $1/se^2$. Standard errors that are clustered by state. Model 1 includes data from 2009 Ed Facts and NAEP data from grades 4 and 8. Model 2 data from 2009 Ed Facts and NWEA data sources from grades 4 and 8. The models are restricted to state or district by grade cells with gap data from both Ed Facts and NAEP/NWEA. Both models also include the proportion of "other" (not shown) items.

Figure 3: Trend in Male-Female Achievement Gaps over Grades



Notes: Estimated trends are taken from 3-level precision weighted HLM models with a non-parametric grade-by-term